# Correlation and Sampling

Gov 51 Section — Week 4

George

Harvard University

February 18, 2026

# Today's Plan

**Part 1: Correlation**

- Quick formula recap
- Matching scatterplots exercise
- Hand calculation (pairs)
- Spot-the-mistake quiz

**Part 2: Sampling**

- SE/MOE/CI formulas
- SE speed drill
- Two full CI problems
- Florida polls in R

Today is mostly **practice**. PS2 is due **Thursday, March 5**.

# Quick Check-In

With a neighbor, answer these in 60 seconds:

1. What does covariance measure?
2. Can covariance be negative? When?
3. What are the *units* of $\text{Cov}(\text{height in inches}, \text{weight in lbs})$?

Warm-up — building on Thursday's lecture.

# Part 1: Correlation

# From Covariance to Correlation

**Problem**: Covariance depends on units, so you can't compare across variables.

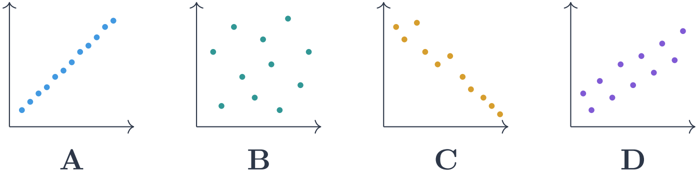
**Fix**: Divide by both standard deviations to get a unitless measure.

$$\textbf{Correlation:} \quad r_{xy} = \frac{\mathrm{Cov}(x, y)}{s_x \cdot s_y} = \frac{s_{xy}}{s_x \, s_y}$$

- Always between $-1$ and $+1$
- Measures **linear** association only
- $r > 0$: positive, $r < 0$: negative, $r = 0$: no linear relationship

# Exercise 1: Match the Scatterplot

Which $r$ goes with which plot? — 2 minutes




**Options**: $r = -0.7$ $r = 0$ $r = 0.4$ $r = 0.9$

Match each plot to its $r$. Discuss with your neighbor.

# Exercise 1: Solution

| Plot | $r$ | Why? |
|---|---|---|
| **A** | $+0.9$ | Tight positive cluster |
| **B** | $0$ | No pattern at all |
| **C** | $-0.7$ | Clear downward trend |
| **D** | $+0.4$ | Upward but scattered |

As $|r| \rightarrow 1$, points cluster tightly around a line.
As $|r| \rightarrow 0$, points form a cloud.

# Exercise 2: Calculate $r$ by Hand

Work in pairs — 5 minutes

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 2 | 3 |
| 2 | 4 | 5 |
| 3 | 6 | 4 |
| 4 | 8 | 8 |
| 5 | 10 | 9 |

**Steps:**

1. Calculate $\bar{x}$ and $\bar{y}$
2. Find deviations: $x_i - \bar{x}$ and $y_i - \bar{y}$
3. Multiply deviations, sum them up
4. Divide by $n-1$ to get $\text{Cov}(x, y)$
5. Calculate $s_x$ and $s_y$
6. Divide: $r = \text{Cov}(x, y)/(s_x \cdot s_y)$

Take 5 minutes. I'll walk around.

# Exercise 2: Solution

| $i$ | $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|---|
| 1 | 2 | 3 | $-4$ | $-2.8$ | 11.2 |
| 2 | 4 | 5 | $-2$ | $-0.8$ | 1.6 |
| 3 | 6 | 4 | 0 | $-1.8$ | 0 |
| 4 | 8 | 8 | 2 | 2.2 | 4.4 |
| 5 | 10 | 9 | 4 | 3.2 | 12.8 |
| $\bar{x} = 6$ | | $\bar{y} = 5.8$ | | | $\sum = 30$ |

$$\mathrm{Cov}(x, y) = \frac{30}{4} = 7.5, \qquad s_x = \sqrt{10} \approx 3.16, \qquad s_y = \sqrt{6.7} \approx 2.59$$

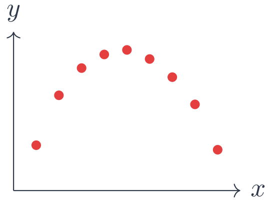
$$r = \frac{7.5}{3.16 \times 2.59} = \frac{7.5}{8.19} \approx 0.916$$

# Verify in R

```r
x <- c(2, 4, 6, 8, 10)
y <- c(3, 5, 4, 8, 9)

cor(x, y)                          # Correlation
## [1] 0.9162

cov(x, y)                          # Covariance
## [1] 7.5

cov(x, y) / (sd(x) * sd(y))       # Manual check
## [1] 0.9162
```

`cor()` does exactly what we did by hand: $\mathrm{Cov}/\left(s_x \cdot s_y\right)$.

# Correlation Measures *Linear* Relationships Only

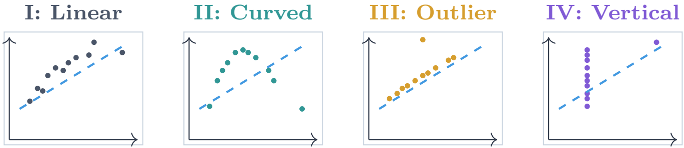



$r \approx 0$

This is a clear pattern but $r \approx 0$ because positive and negative deviations cancel.

$r = 0$ does NOT mean "no relationship." It means no *linear* relationship.

# Anscombe's Quartet: Same $r$, Different Data




All four: **same** $\bar{x}$, same $\bar{y}$, same $s_x$, same $s_y$, same $r = 0.82$, same regression line.

Summary statistics can hide important patterns. **Always plot your data.**

# Exercise 3: Spot the Mistake

Which statements are WRONG? — 2 minutes, then discuss

Three students said the following about correlation. Which are correct?

1. "The correlation between study hours and GPA is $r = 1.3$, which means studying is strongly related to grades."
2. "Ice cream sales and drowning deaths have $r = 0.85$, so eating ice cream causes drowning."
3. "The correlation between $x$ and $y$ is $r = 0$, so there is absolutely no relationship between them."

All three are wrong! Can you explain why for each?

# Exercise 3: Why They're Wrong

1. **$r = 1.3$ is impossible.** Correlation is always between $-1$ and $+1$. Someone made a calculation error.
2. **Correlation $\neq$ causation.** Both are driven by a third variable (summer heat). High $r$ means they move together, not that one causes the other.
3. **$r = 0$ means no *linear* relationship.** There could be a curved relationship (like the parabola we just saw).

Three common mistakes: impossible values,
causal language, and forgetting the "linear" part.

# Exercise 4: Spurious Correlations

2 minutes in pairs

Come up with **two examples** of variables that are correlated but clearly NOT causally related.

**Hint**: Think about what *third variable* might drive both.

Share your best example with the class.

Classic examples: shoe size & reading level (age), Nicholas Cage films & pool drownings...

# Part 1 Summary

**Covariance:** $s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$

$\big\downarrow$ standardize

**Correlation:** $r_{xy} = \frac{s_{xy}}{s_x \, s_y}$

- Correlation: direction **and** strength ($-1$ to $+1$, unitless)
- Measures **linear** relationships only — always plot your data
- Correlation $\neq$ causation

# Part 2: Sampling and Uncertainty

# Formulas You Need

**Standard Error:** $\text{SE} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

**Margin of Error:** $\text{MOE} = 1.96 \times \text{SE}$

**95% CI:** $\hat{p} \pm 1.96 \times \text{SE}$

- SE = how much $\hat{p}$ would vary across repeated samples (not SD!)
- The 1.96 comes from the normal distribution (95% coverage)

# What Does a 95% CI Actually Mean?

**Correct**: If we repeated this procedure many times, 95% of the resulting intervals would contain the true value.

**Wrong**: "There is a 95% chance the true value is in this interval." The true value is fixed.

The confidence is in the *procedure*, not in any single interval.

# Exercise 5: SE Speed Drill

4 minutes — calculate all three

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

| | Scenario | $\hat{p}$ | $n$ |
|---|---|---|---|
| (a) | A poll finds 54% support a candidate | 0.54 | 900 |
| (b) | 62% of students prefer coffee over tea | 0.62 | 200 |
| (c) | 50% of coins land heads | 0.50 | 10,000 |

Calculate the SE for each. Which has the smallest SE? Why?

# Exercise 5: Solutions

(a) $\text{SE} = \sqrt{\frac{0.54\times0.46}{900}} = \sqrt{0.000276} \approx \mathbf{0.0166}$

(b) $\text{SE} = \sqrt{\frac{0.62\times0.38}{200}} = \sqrt{0.001178} \approx \mathbf{0.0343}$

(c) $\text{SE} = \sqrt{\frac{0.50\times0.50}{10{,}000}} = \sqrt{0.000025} = \mathbf{0.0050}$

(c) has the smallest SE because $n$ is huge.
Bigger samples $\rightarrow$ more precise estimates.

# Exercise 6: What Happens to MOE?

Think, then share — 3 minutes

For each scenario, predict: does the MOE get **bigger**, **smaller**, or **stay the same**?

1. You **double** the sample size (from 400 to 800)
2. You change from polling about a 50–50 race to a 90–10 race
3. You want to **halve** the MOE. How many times bigger does $n$ need to be?
4. A pollster switches from $n = 1{,}000$ to $n = 1{,}500$. Roughly how much does MOE change?

# Exercise 6: Answers

1. **Smaller** — but only by factor of $\sqrt{2} \approx 1.41$, not by half
2. **Smaller** — $\hat{p}(1-\hat{p})$ is maximized at $\hat{p} = 0.5$, so moving away from 50–50 reduces SE
3. **4 times** bigger — the square root rule. To halve MOE, you need $4\times$ the sample.
4. From $\pm 3.2$ pts to $\pm 2.6$ pts (about 20% smaller). Diminishing returns!

The square root rule: $4\times$ the sample to halve the MOE.
This is why most polls use 1,000–1,500 respondents.

# Exercise 7: Full CI Calculation

5 minutes — work through all three steps in pairs

A poll of **1,200 likely voters** finds that **47%** support a ballot measure.

1. Calculate the **standard error**
2. Calculate the **margin of error**
3. Construct the **95% confidence interval**

**Bonus**: Based on your CI, is this race a toss-up? How do you know?

Take 5 minutes. Work in pairs.

# Exercise 7: Solution

1. $\text{SE} = \sqrt{\frac{0.47 \times 0.53}{1200}} = \sqrt{0.000208} \approx 0.0144$
2. $\text{MOE} = 1.96 \times 0.0144 \approx 0.028$ (about 2.8 percentage points)
3. $\text{CI} = 0.47 \pm 0.028 = [0.442,\ 0.498]$ or $[44.2\%,\ 49.8\%]$

**Bonus**: 50% is just barely *outside* the CI (upper bound is 49.8%), so the candidate is slightly behind — but it's very close to a toss-up.

# Exercise 8: Another CI Problem

3 minutes — on your own this time

An exit poll of **2,500 voters** finds that **52%** voted for Candidate A.

1. Calculate the SE
2. Construct the 95% CI
3. Can we confidently say Candidate A won?

You should be getting faster at this!

# Exercise 8: Solution

1. $\text{SE} = \sqrt{\dfrac{0.52 \times 0.48}{2500}} = \sqrt{0.0000998} \approx 0.0100$
2. $\text{CI} = 0.52 \pm 1.96 \times 0.0100 = 0.52 \pm 0.020 = [0.500,\ 0.540]$
3. 50% is right at the lower boundary. We **cannot** confidently say A won — the race is too close to call.

This is why election night is stressful! A 2-point lead with $n = 2{,}500$ is within the margin of error.

# R Application: The 2008 Election

# The Polling Data

We have 1,333 state-level polls from the 2008 Obama–McCain election.

```r
library(tidyverse)

polls <- read_csv("polls08.csv")
nrow(polls)
## [1] 1333
```

```r
# How many states?
n_distinct(polls$state)
## [1] 51
```

This is the same data you'll use in PS2.

# Florida: A Swing State

```r
florida <- polls |> filter(state == "FL")

nrow(florida)            # 73 polls!
mean(florida$Obama)      # Average Obama support
sd(florida$Obama)        # How much do polls vary?
```

**Why do polls disagree?**

- Sampling error (random variation)
- Different timing (opinions shift)
- Different methods (phone vs. online)

Aggregating polls reduces sampling noise.

# Calculate SE and CI for Florida

```r
p_hat <- mean(florida$Obama) / 100
n <- 1000  # Typical poll sample size

se <- sqrt(p_hat * (1 - p_hat) / n)
ci_lower <- p_hat - 1.96 * se
ci_upper <- p_hat + 1.96 * se

cat("SE:", round(se, 4), "\n")
cat("95% CI:", round(ci_lower, 3), "to",
    round(ci_upper, 3), "\n")
```

Obama actually won Florida with about 51%. Does your CI contain this value?

# Wrapping Up

# Key Formulas Reference Card

| Concept | Formula | R function |
|---|---|---|
| Correlation | $r = \frac{\mathrm{Cov}(x, y)}{s_x \cdot s_y}$ | `cor(x, y)` |
| Standard Error | $\mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ | (calculate manually) |
| Margin of Error | $\mathrm{MOE} = 1.96 \times \mathrm{SE}$ | (calculate manually) |
| 95% CI | $\hat{p} \pm 1.96 \times \mathrm{SE}$ | (calculate manually) |

# For the Exam

You should be able to:

1. **Calculate and interpret correlation** — by hand and in R
2. **Explain why $r$ can be misleading** — Anscombe's quartet
3. **Calculate SE, MOE, and 95% CI** from poll results
4. **Correctly interpret a confidence interval** — the frequentist way
5. **Explain the square root rule** — $4\times$ sample to halve MOE

PS2 covers all of these topics. Use it as exam practice!
Due: **Thursday, March 5** at 11:59pm.

**Correlation measures linear association ($-1$ to $+1$). Standard error measures sampling uncertainty. Always plot your data.**

Questions?